

Having thus described our invention, we now claim:

Sub
a1

1 1. A method for administration and replication of a database, comprising the
2 steps of:
3 providing a database management system with a built-in random
4 sampling facility integrated into said database management system; and,
5 executing said random sampling facility from within the database
6 management system to perform a replication operation on said database.

1 2. The method as set forth in claim 1, further comprising the steps of:
2 defining a database record sample size S;
3 randomly sampling S records of the database using said random sampling
4 facility;
5 storing statistics for each of said S records, wherein said statistics include
6 a record key for each record; and,
7 producing an extrapolated replication partition analysis based on said
8 statistics.

1 3. The method as set forth in claim 2, wherein the step of defining said
2 sample size S includes:
3 defining a default sample size;
4 selectively receiving a desired sample size; and,

Sub
a1

5 setting said sample size S as said default sample size when the desired
6 sample size is not selectively received, and setting said sample size S as said desired sample size
7 when the desired sample size is selectively received.

1 4. The method as set forth in claim 1, further comprising the steps of:
2 defining a database record sample size S;
3 randomly sampling S records of the database using said random sampling
4 facility;
5 storing statistics for each of said S records, wherein said statistics include
6 a record key for each record; and,
7 producing a partial replication partition analysis based on said statistics.

1 5. The method as set forth in claim 4, wherein the step of defining said
2 sample size S includes:
3 defining a default sample size;
4 selectively receiving a desired sample size; and,
5 setting said sample size S as said default sample size when the desired
6 sample size is not selectively received, and setting said sample size S as said desired sample size
7 when the desired sample size is selectively received.

1 6. A method for database administration and replication, comprising the
2 steps of:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 1040 1

1 7. The method as set forth in claim 6, wherein the step of selecting said
2 default sample size value D further includes the steps of:
3 generating a table of S number pairs (Y_j, I_j) , $j=1,2,...,S$, wherein all Y and
4 all I are initially set to zero;
5 initializing a reservoir of records to an empty +state;
6 setting an index M to said reservoir equal to zero;
7 generating a sequence of N non-repeating random numbers $U_1, U_2, ..., U_N$,

Sub
(a)

8 $0 \leq U \leq 1$, wherein N is the number of records in the database; and,

9 performing additional steps for each random number U_k generated,

10 $k=1,2,\dots,N$, the additional steps including:

11 skipping the next record in the database if U_k is less than the

12 smallest value of Y in said table of number pairs; and,

13 updating the table if a Y less than U_k exists by performing

14 further steps including:

15 setting M equal to its current value plus one;

16 replacing the smallest Y in the table with U_k ;

17 setting the I value paired with the smallest Y equal

18 to M; and,

19 storing all or part of the next record of the

20 database in said reservoir of stored records, wherein the current value of

21 M is a reservoir index to said stored record.

1 8. The method as set forth in claim 7, wherein the step of updating the table

2 further includes the step of:

3 arranging the table in a heap with respect to Y.

1 9. The method as set forth in claim 6, further comprising the step of:

2 sorting said stored statistics by key prior to producing said partition

3 analysis.

Sub
a1

1 10. The method as set forth in claim 9, further comprising the steps of:
2 accessing all database records in an arbitrary sequence;
iteratively filling all of said partitions except the last said partition with
said accessed records to a maximum byte count; and,
storing remaining accessed records in the last of said partitions.

1 11. The method as set forth in claim 6, wherein the step of storing statistics
2 includes storing said statistics in a memory.

1 12. The method as set forth in claim 11, wherein the step of storing statistics
2 includes storing said statistics in said memory in a compressed format.

1 13. The method as set forth in claim 6, wherein the step of producing at least
2 one of said partition analyses includes the step of defining multiple partition boundaries.

1 14. The method as set forth in claim 6, wherein the step of sampling said S
2 records includes randomly sampling the S records utilizing dataspace including:
3 at least one index datasource;
4 at least one key datasource; and,
5 at least one statistics datasource.

09897803-070201

Sub
a1

1 15. A database management system (DBMS) for managing an associated
2 database, the DBMS comprising:
3 random sampling facility integrated with the database management
4 system;
5 first database analysis tools using said integrated random sampling
6 facility for generating extrapolated reports on database content;
7 second database analysis tools using said integrated random sampling
8 facility for generating extrapolated reports on database size; and,
9 database replication tools adapted to execute at least one of a complete
10 replication having output partition sizes determined by extrapolating a random sample of said
11 database, and a partial replication in which the data stored in the partial replication comprises a
12 random sample of said database.

1 16. The database management system of claim 15 further comprising:
2 a pre-configured number S defining a default sample size;
3 a means for selectively receiving a particular number defining a desired
4 sample size and setting said number S equal to said particular number;
5 a means for randomly sampling S records of the database using said
6 random sampling facility;
7 a means for storing statistics for each of said S records, wherein said
8 statistics include a record key for each record; and,
9 a means for producing at least one of:

Sw
a1

an extrapolated database content analysis based on said statistics;
an extrapolated partition analysis based on said statistics; and,
a partial partition analysis based on said statistics.

17. The database management system of claim 16, further comprising:
a means for sorting said stored statistics by key prior to producing at least
one of said analyses.

18. The database management system of claim 16, wherein said means for
randomly sampling S records further comprises:

a means for generating a table of S number pairs (Y_j, I_j) , $j=1,2,...,S$,
wherein all Y and all I are initially zero;

a means for initializing a reservoir of records to an empty state;

a means for setting an index M to said reservoir equal to zero;

a means for generating a sequence of N non-repeating random numbers
 $U_1, U_2, ..., U_N$, $0 \leq U \leq 1$, wherein N is the number of records in the database; and,

a means, for each random number U_k generated, $k=1,2,...,N$, comprising:

a means to skip the next record in said database if U_k is
less than the smallest value of Y in said table of number pairs; and,

a means to update the table if a Y less than U_k exists,
comprising:

a means to set M equal to its current value plus one;

Sub
a1

15

a means to replace the smallest Y in the table with U_k ;

16

a means to set the I value paired with the smallest Y equal

17

to M; and,

18

a means to store all or part of the next record of said

19

database in said reservoir of stored records, wherein the current value of

20

M is a reservoir index to said stored record.

1

19. The database management system of claim 18 wherein the means to

2

update the table further comprises:

3

a means to arrange the table in a heap with respect to Y.

1

20. The database management system of claim 18, wherein said means for

2

storing statistics comprises a means for storing said statistics in memory.

1

21. The database management system of claim 20, further comprising a

2

means for sorting said stored statistics by key prior to producing at least one of said analyses.

1

22. The database management system of claim 21, wherein said partition

2

analyses include analyses of multiple partition boundaries.

1

23. The database management system of claim 22, further comprising:

2

a means for accessing all database records in an arbitrary sequence;

Sub
a1

3 a means for iteratively filling all of said partitions except the last with said
4 accessed records to a maximum byte count; and,
5 a means for storing remaining accessed records in the last of said
6 partitions.

1 24. The database management system of claim 16, further comprising:
2 a means for utilizing at least one index dataspace;
3 a means for utilizing at least one key dataspace; and,
4 a means for utilizing at least one statistics dataspace.

1020/0" E0B/6860